Trustworthy to Trusted: Building Data Consumer Confidence in Your Data Assets

Jay Hollingsworth

CTO

Energistics

Abstract

Knowledge workers need data they can trust as early in their processes as possible to make decisions that improve operational efficiency, enable worker safety and minimize costs.

As our previous papers on trustworthiness have shown, data do not have to be of uniformly HIGH quality to be useful, they have to be of APPROPRIATE quality for the task at hand. To judge whether the data are appropriate or not requires the availability of additional information - "metadata" - so the user can decide what to use and what to leave behind.

This paper builds on the previous papers to suggest methods for extracting this metadata directly from acquisition and interpretation systems and how it could best be conveyed.

Introduction

Knowledge workers know the quality of their work or the decisions they make will be adversely affected by bad or inappropriate data. These knowledge workers–geoscientists and engineers, data scientists, analysts, planners, managers, etc.–are individually uncomfortable using data they do not trust. They will spend as much time as they feel is necessary examining and “correcting” the data they need to use to perform their work. Factors influencing how long this process will take include their individual feelings about the sources of information, prior bad experiences with use of improper data, and the riskiness and visibility of the result of their work product.

Establishing or recovering a level of trust in the official sources of enterprise data is a key – though often unwritten – objective of corporate data management efforts. Many internal IT projects use the word “trusted” in their titles. Some commercial applications which focus on creating trusted data use a variant of “trust” in their product name. These are all recognition of the value of trust in the data chain.

In previous papers[1, 2] we established the fact that trust is a different concept from data quality, and that in fact people trust the use of poor quality data so long as they explicitly understand the level of trustworthiness and can assess whether their use of imperfect data is appropriate for their task at hand; they simply take the uncertainty into account. We also enumerated the dimensions of trust which exist in E&P data and showed that each could be explicitly measured and scored.

In the second paper we examined more closely those dimensions which relate directly to classical uncertainty measures and considered the impact on data management practices in supporting data scientists and other knowledge workers as they consider which data to use for a given task (or as they try to improve the quality of the data at hand).

In this paper we look at the practicalities of obtaining the descriptive information–metadata–that satisfy the need to understand levels of trustworthiness and propose patterns for moving this information alog with the data it describes.

[1] From Data Transfer to Data Assurance: Trusted Data is the Key for Good Decisions, 22nd PNEC Conference & Exhibition, Jay Hollingsworth *et al.*

[2] The Life Cycle of Trusted Data: From Acquisition to Persistence... Or Not, 23rd PNEC Conference & Exhibition, Jay Hollingsworth *et al.*

Dimensions of Trust

In our prior work we elected to use definitions of trustworthiness put forth by IBM[2] in 2011 and added to them. We use this framework to show that a data transfer standard could be used to convey the components of trustworthiness from one system to another so a user could establish their own level of trust in a data source. There are additional details – facets – along the dimensions which we will not repeat here.

There are three dimensions of trust:

1. Trust in the Data Source
   The qualitative reputation and a quantification of the reliability of a source of data inherits to the data retrieved from that source.

2. Data Lineage
   Data lineage creates trust in a data by tracing its provenance from its origin through the history of its processing.

3. Data Security
   This dimension encompasses trust in the security of the underlying systems and (in the case of real-time data) security of the transmission mechanism between the sensor and the user.

Some authors have created schemes to try to assign scores which relate to the trustworthiness of a particular piece of data or of data sources. Most of this work relates to information of a more qualitative nature – trust in government or financial institutions, for example. These scores are derived from assigning values to the various facets which exist within the three dimensions above (see the referenced papers for more details).

Creating Trusted Datasets

Knowledge workers – geoscientists, engineers, planners, data scientists, anyone who works with data – are not comfortable performing their jobs with data they do not trust. Our prior work showed that they will not consider that they have begun to do their work until they have a set of data they can trust. They will take whatever amount of time is required to create that trusted dataset.

This data preparation and cleanup time – called data wrangling or data munging – is the time commonly called time spent looking for data. It reportedly takes from 60% to 80% of a knowledge worker's time.

This time is not wasted, because during the process of data munging the user will develop insights from the kind of close examination of the data and relationships among data needed to do the cleanup required.

[3] The Third International Conference on Advances in Databases, Knowledge, and Data Applications, Trusted Data in IBM's Master Data Management, Przemyslaw Pawluk *et al.*

Trusted Sources of Data

We showed in the previous work that data consumers will clearly spend less time wrangling data from sources they trust. This sense of trust comes from an awareness of the degree to which the three trustworthiness dimensions summarized above are satisfied by their sources of data for the task to which it will be put.

This sense of trustworthiness for a particular purpose is dependent on the task at hand within a workflow, in addition to a characteristics of the data available and the user's prior experiences with that source.

Trusted Data Is Not Necessarily Quality Data

Trusted data is not the same thing at all as "quality" data.

For example, if a Production Accountant needs to allocate and report monthly production volumes from a group of wells they will be much more concerned about the accuracy of the well tests and the monthly flow readings from each individual well than they would be about the precision of the wells' surface locations. If they previously had problems with the quality of the monthly readings they will spend more time researching the source and handling of that than they would the other information they needed. The well locations – while important in other contexts and may be reported along with the allocated volumes – would not be a concern since they aren't needed for the allocation computations; the location data would be sufficiently trustworthy for this particular task, even if there was some doubt as to their accuracy.

In this example, the Accountant will research the lineage and handling of the volumes data because of prior experience even though it may be the highest quality from recently-calibrated instruments and will not focus on data of admittedly lower quality because it is not critical to the task at hand.

Measurements

It is intuitive that information entered into any kind of system manually will be less trustworthy than values measured directly.

Modern measurements concern themselves with their accuracy and uncertainty, and any measurement apparatus will have some kind of traceability back to that standard, even if that path is no longer known.

Trust in a measurement also depends on the user of the data being aware of the process which has been followed in acquiring a value and deciding whether that process has resulted in a value suitable to the use at hand.

The Process of Measurement

The process of measurement in E&P has several steps, each of which introduces its own uncertainties and possibilities for error. Data managers need to decide now many of these steps they will present to their users to establish and retain their trust in the official source of truth:

- Identification of a desired measurement, followed by

- Realizing that it can't actually be measured, but that something related can be
- Selection and acquisition of a sensor to measure that related thing
- Placement of the sensor
- Calibration of the sensor
- Taking of a reading
- Conversion of the analog reading to a digital value
- Calculation of the desired measurement from the actual one
- Relaying the measurement to the end user and finally
- Managing the measurement results to ensure their enduring integrity

In the previous work we examined the measurement process and concluded that trust in measured values follows the same pattern as all other kinds of information – for a user to trust any particular piece of information they need access to the contextual facets defined within the dimensions mentioned above, and the more trust they have in the immediate source of the data the less time they will spend examining that context.

The challenge for data managers who oversee those direct-to-user sources of information is to build up this sense of trust so their knowledge worker colleagues spend as little time as possible in data wrangling and as much as possible in value-added analytics. This sense of trust is built up through experience and easy availability of as much lineage and contextual information as possible.

Building Consumer Confidence

The process of building user confidence in an official source of data is a continuous one – a one-time investment in a data quality or master data management program with a consultant will not create a permanent aura of trustworthiness around a particular data warehouse. Especially where users have had negative consequences from prior efforts using the same source or from the same staff these projects may not improve the reputation at all.

Users will have to make up their own minds based on continuous availability of contextual information as described above. Where did the data come from originally? Has any calculation been made (corrections, adjustments, depth shifts, etc.) to the value I have available to me? Who has had access to the information who could have accidentally replaced good with bad numbers?

The contextual information needed to create or restore user confidence in data assets is metadata.

Metadata

Metadata is often defined as data about data, but that simple definition doesn't really convey the extent of what is meant. A better definition of metadata comes from an older (2004) edition of "Understanding Metadata" from the National Information Standards Organization:

> *[Metadata is] structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. Metadata is often called data about data or information about information.*

This definition covers the different aspects that upstream data managers need to consider. Metadata is **structured** and it **describes**, **explains**, and **locates** data assets and makes them easier to **find** and **use**.

The original and simplest forms of metadata are the (now) old-fashioned library card catalog. Each kind of book in a library had a matching index card which described the book; these cards were stored in long drawers. There would be multiple sets of these card-filled drawers in large cabinets – some organized by author, others by title, still others arranged by topic or other aspects of the books contents.

The oldest known card catalog to survive in part is the Assyrian Royal Library of Ashurbanipal dating from around 630 BCE, most of which is now in the British Museum. As libraries, museums and other formal inventories grew over time their keepers created formalisms to describe the items and their contents.

With the advent of computers the paper catalogs became computer databases which could exchange records of these information assets and search through them. It was when people stopped thinking about books as objects with content in them and began thinning of them as data assets that the formal notion of "metadata" evolved.

Categorizing Metadata

There are numerous schemes for categorizing metadata, but the concept developed in this paper is less concerned with what is contained in the metadata – which may be somewhat specific to certain kinds of information assets – and is more concerned with how the metadata is connected to the asset, how it is created and how it can be used to build consumer confidence in the data asset itself.

To consider how metadata could be connected to its object, we need to cagorize metadata as follows: independent metadata, referenced metadata and intrinsic metadata.

A book metaphor will help explain the concepts needed.

Independent Metadata

Libraries have compiled metadata since at least Assyrian time. This metadata was the ancestor of the physical card catalog or later computerized library databases. There is an entry in the catalog for each type of book the library holds, and possibly and entry for each individual copy of a book.

The book as published would be unaware that a metadata record exists describing it. Most libraries would write a (Dewey Decimal or LC in the U.S.) code on the spines of the books so they can be sorted on the shelves and more easily located, but without these alterations the books would carry no indication that someone had cataloged them.

The metadata is completely independent of the books as published.

The same is true of other kinds of catalogs–museums and other collections maintain catalogs but the items cataloged normally have no reference back to that metadata.

Referenced Metadata

Modern books as published normally have a standard book number (ISBN) printed in the book on the reverse of the title page and next to the barcode on the back. Some books will have a different number on the barcode itself. These numbers are references to the descriptive metadata about the book.

The ISBN number can be used to find out information about the nature of the book itself; if there is another number on the barcode it is a reference to the inventory and sales system of the publisher.

In this case, the object being described–a book–carries an identifier which references metadata about it.

If the consumer of the object has access to the catalog where the object's metadata is stored he or she can inspect and use that metadata.

It's worth noting that the metadata used as examples of referenced metadata is actually for a class of objects, not for the location or description of a single occurrence of that object.

Intrinsic Metadata

The most powerful connection between an object and its metadata is for the object to carry its own metadata.

Electronic books are able to carry any kind of standard metadata and they normally hold who created them and when, whether the document is editable or not, and many other items.

In oil & gas there are many data artifacts which include this kind of information, including the SEG standards, the IOGP positioning standards and all the Energistics transfer standards.

This metadata is often extracted from the objects of interest and made available in a separate catalog for ease in browsing or searching. But the key benefit of this metadata is that it cannot be out of sync with the object – assuming the metadata was correct in the creation of the object itself, if the data in the catalog develops a problem one can just re-import the intrinsic metadata from the object.

Sources of Metadata

There are many sources of metadata.

The simplest – and most error-prone – is manual entry of metadata into some kind of catalog. The best sources of metadata are an extraction from the object's intrinsic metadata, metadata carried in a scientific interpretation application and metadata which flows from a sensor together with its measurements. There are additional sources as well – public information, other curated databases inside a company, process historians and other operational data systems, control systems, maintenance management applications – but these sources require additional work to match up the metadata with the data object being described; that match-up process is another potential source of error.

The best sources of metadata mentioned above also have the advantage of offering the possibility of automatically populating metadata stores since there is no need to match the object with its metadata after the fact – the metadata arrived with the data object or sensor reading.

Regaining User Trust

The key to developing user confidence in a source of data – or to regaining that trust if it has been lost – is in making this metadata available and easily visible to an end user. There may be some few users who will habitually check the heritage of every value, but even the most cynical after a while will understand the source of their information and will know the kinds of uses to which it can be put.

Most users will not go to this kind of extreme; they are too busy or believe they have too much time pressure. The key for them – and the starting point for the more doubting user – is simply to carry metadata and make it trivially available. This is the most important point of our thesis:

Simply curating and making metadata easily available alongside data is the key to trust

The final aspect of building and maintaining consumer trust is for there to be practical commercial standards for conveying any arbitrary structured metadata along with data objects and sensor readings.

Data Transfer Standard Metadata

The latest generation of Energistics standards – WITSML v2.0, PRODML v2.0 and RESQML v2.0.1 – all fully support the metadata needed to describe these aspects.

Data Assurance Object

The DataAssurance object is common across all of the Energistics standards.

It declares whether each bit of data contained in a related standard object fails a pre-defined corporate data policy. To save bandwidth the Data Assurance object isn't normally carried for data which conforms to the policy, since that is expected to be the norm.

A Data Policy is:

> One or more rules that must be applied to assess the data. Conformance to a policy would classify the data as Trusted

So a Policy is just a named set of rules which trustworthy data should follow. An exception to a policy is flagged as a failure along with the identities of the rules within the policy which failed.

The rules may be determined by the values of the data itself, or may be defined externally.

For example, a company could have a policy called the "Well Location Adequacy Policy." Within this policy could be three rules:

1. Every well location must be re-surveyed with 30 days of completion of the well
2. Every well location must be given in WGS84 as a latitude, longitude pair
3. Every well location must be in a map projection as a northing, easting pair

In a case where all three rules are satisfied there would be no Data Assurance object. In the event that it has been two months and the location hasn't been re-surveyed yet, any transfer of the well location data would carry a Data Assurance object which would note that the data fails the corporate Well Location Adequacy Policy because a new survey is not present.

The Data Assurance Object is in the shared portion of the new suite of Energistics standards, so it is available to be used by all of them and for workflows that require a combination of Energistics standards.

Activity Model

Another new feature common to all the standards is the Activity model.

The Activity objects carry knowledge that an activity was performed (and who did it and when using what software) which may have consumed data, may have used certain parameters and may have produced some data. These could be real E&P field activities like perforating or calibrating a meter, or could be a calculation activity like starting up a reservoir simulator.

For example I could have an activity which uses Archie's equation, the neutron porosity channel called NPHI1, the resistivity channel called LLD2 and cementation and saturation factors of 2.0 to compute the water saturation curve WS3.

The Activity objects can be carried along with any WITSML object, just like Data Assurance or can be transported independently.

Trusted Data Via Data Assurance in Energistics Standards

The key point about the Data Assurance process embodied in the new generation of Energistics' standards is that the data – flawed or not – is always transferred. There is just some extra descriptive metadata carried along as needed to describe the processes used to create or treat the data.

It should be clear that the Activity model and the Data Assurance object taken together satisfy all the facets of trust defined before and could describe the aspects of uncertainty presented here.

The Data Lineage and Data Security facets are covered by use of the activity model. The Trust of Data Sources group is satisfied by defining appropriate rules within policies.